
Monograph:
Gutacker, Nils orcid.org/0000-0002-2833-0621 and Street, Andrew David orcid.org/0000-0002-2540-0364 (2015) Multidimensional performance assessment using dominance criteria. Discussion Paper. CHE Research Paper . Centre for Health Economics, University of York , York, UK.

ESHCRU
Economics of Social and Health Care Research Unit

CHE
Centre For Health Economics

University of York

Multidimensional Performance Assessment Using Dominance Criteria

Nils Gutacker, Andrew Street

CHE Research Paper 115

Multidimensional performance assessment using dominance criteria

Nils Gutacker
Andrew Street

Centre for Health Economics, University of York, UK

September 2015

Background to series

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Acknowledgements

We are grateful for comments and suggestions from Stirling Bryan, Chris Bojke, Richard Cookson, Bernard van den Berg, Pedro Saramago, Miqdad Asaria and those received during the HESG Summer 2012 meeting (Oxford), the 2014 iHEA and ECHE conference (Dublin), and the 34th Spanish Health Economics Association conference (Pamplona). This research was funded by the Department of Health in England under the Policy Research Unit in the Economics of Health and Social Care Systems (Ref 103/0001). The views expressed are those of the authors and may not reflect those of the Department of Health.

Further copies

Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/ Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321405 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/che

Abstract

Public sector organisations pursue multiple objectives and serve a number of stakeholders. But stakeholders are rarely explicit about the valuations they attach to different objectives, nor are these valuations likely to be identical. This complicates the assessment of their performance because no single set of weights can be legitimately chosen by regulators to aggregate outputs into unidimensional composite scores. We propose the use of dominance criteria in a multidimensional performance assessment framework to identify best practice and poor performance under relatively weak assumptions about stakeholders' preferences. We estimate multivariate multilevel models to study providers of hip replacement surgery in the English NHS with respect to their performance in terms of length of stay, readmission rates, post-operative patient-reported health status and waiting time. We find substantial correlation between objectives and demonstrate that ignoring the correlation can lead to incorrect assessments of performance.

Keywords: Performance assessment, provider classification, multidimensional, multilevel modelling

1. Introduction

Variation in healthcare quality and costs are well documented (Wennberg and Gittelsohn 1973; Keeler 1990; Busse et al. 2008; Bernal-Delgado et al. 2015) and may arise when providers enjoy discretion over how their services are organised and provided (Arrow 1963). Regulators, who are charged with overseeing the provision of care, are concerned about variation if it is not caused by differences in healthcare needs or patient preferences as it may signal inequity, inefficiency or unsafe care. To address this, many healthcare systems have implemented routine benchmarking (or 'profiling') of healthcare providers to identify comparative performance levels. This might help single out 'positive deviants' (Bradley et al. 2009; Berwick 2008; Lawton et al. 2014), or exemplars of best practice, that can be studied further or rewarded as part of a pay-for-performance scheme. At the other extreme, poor performers might be subject to penalties for falling short of their peers or to interventional actions by regulators.

Healthcare providers share two important features with other public sector organisations that complicate the assessment of their performance (Dixit 2002; Besley and Ghatak 2003; Propper and Wilson 2012). First, they lack a single overarching objective, such as profit, against which their performance can be assessed. Instead, they pursue multiple, sometimes conflicting, objectives and this requires the regulator to measure and incentivise achievements along a range of performance dimensions. These achievements are typically non-commensurate and include different aspects of performance reflecting resource use, clinical effectiveness, and other dimensions of quality such as accessibility (Smith 2002; Goddard and Jacobs 2009; Porter 2010; Devlin and Sussex 2011). Second, providers typically serve several stakeholders (e.g. patients, purchasers of care, and politicians) and the values these stakeholders attach to objectives are often not known to the regulator[1], but are unlikely to be identical (Smith 2002; Propper and Wilson 2012); see Devlin and Sussex (2011) for examples from healthcare and the wider public sector.

The lack of a set of common, explicit valuations for individual performance dimensions makes it difficult to construct a single, unidimensional performance measure. If valuations were known and common across stakeholders, it would be possible to aggregate multiple performance scores into unidimensional composite scores. Such measures are attractive as they allow a complete and transitive ranking of providers, facilitate the presentation and dissemination of performance information to stakeholders, and offer a simple means to adjust rewards in a pay-for-performance framework (Dowd et al. 2014). But without such knowledge, there is no guidance on how to aggregate achievements appropriately.

The empirical literature has addressed this problem in different ways: Some studies restricted their assessment of provider performance to those performance dimensions for which explicit valuations have been expressed. Examples include Timbie et al. (2008), Timbie and Normand (2008) and Karnon et al. (2013), all of which translate hospital mortality estimates into monetary units using the expressed valuation of a statistical life. The obvious shortcoming of this approach is that performance dimensions which lack explicit valuations (e.g. waiting times, patient satisfaction, or emergency re-admission rates[2]) are necessarily omitted from the analysis. Their omission may lead to tunnel vision, whereby providers

[1] One could estimate the preferences of individual stakeholders or groups thereof by means of elicitation or through the study of revealed preferences (Ryan et al. 2001). However, this would likely be a very difficult and costly undertaking and is therefore rarely done in practice.

[2] It may be possible to translate achievements on some objectives, e.g. emergency readmissions rates or other measures of health outcomes, into quality-adjusted life years (QALYs) by means of modelling (Timbie et al. 2009; Appleby et al. 2013; Coronini-Cronberg et al. 2013). A monetary valuation of QALYs has been expressed in the English NHS and elsewhere. However, the data requirements are substantial and the statistical uncertainty introduced through modelling is likely to further compound the problem of differentiating between true performance signal and noise.

concentrate their efforts on performance dimensions with explicit valuations at the expense of other dimensions (Holmström and Milgrom 1991; Goddard et al. 2000).

Alternatively, analysts often either choose a set of weights, implement pre-defined scoring algorithms such as equal weighting, or derive weights from the data using approaches based on item response theory (Landrum et al. 2000; Landrum et al. 2003; Daniels and Normand 2006; Teixeira-Pinto and Normand 2008), data envelopment analysis (Dowd et al. 2014), and more ad-hoc econometric specifications (Chua et al. 2010). However, such practice conflicts with one of the key tenets of economic welfare theory, namely that the stakeholders are the only legitimate judges of their own preferences and that, ultimately, responsibility for specifying valuations for performance dimensions should rest with the relevant stakeholders (Smith and Street 2005). There is no guarantee that weights imposed by analysts, however these are arrived at, match the preferences of all stakeholders. Consequently, organisations being assessed might legitimately question the validity of the generated index.

There is an alternative way to address the problem of determining appropriate weights. Multidimensional performance assessment circumvents the issue by analysing performance against each achievement individually and then combining the results into an overall performance profile. In doing so, it makes explicit how healthcare providers perform on each performance dimension and how these dimensions correlate. The multidimensional approach has enjoyed increasing popularity in the health economic literature: Hall and Hamilton (2004) assess the performance of surgeons in terms of 30-day mortality and morbidity using a Bayesian hierarchical bivariate probit model. Hauck and Street (2006) use multivariate multilevel models to study the performance of health authorities across 13 performance indicators. Gutacker et al. (2013) study hospital performance with respect to five health dimensions and compare their results to those based on a composite measure. Portrait et al. (2015) compare Dutch Diabetes care groups in terms of costs and a broad range of quality indicators, whereas Häkkinen et al. (2014), Kruse and Christensen (2013) and Street et al. (2014) study the performance of hospitals in terms of costs and a single measure of patient health outcome for different conditions.

But multidimensional performance assessment is not a panacea for the problem of judging performance across multiple objectives. A multidimensional performance profile does not permit ranking of hospitals or comparison to some performance standard. Hence it remains unclear which providers excel or perform poorly across multiple performance dimensions. This constitutes a major limitation of the multidimensional approach for practical purposes, and one that we seek to overcome in this study. More specifically, we propose the use of dominance criteria to judge hospital performance against a multidimensional benchmark. The concept of dominance has the attractive feature that it allows comparison of multidimensional performance profiles against benchmarks under relatively weak assumptions about stakeholders' utility functions. Indeed, the only requirement is that the regulator can judge whether the marginal utility of an achievement is positive or negative and that this qualitative judgement applies to all stakeholders. We believe this to be a reasonable pre-requisite in most contexts.

We apply our approach to data on providers of hip replacement surgery in the English NHS during the period April 2009 to March 2012. Performance is assessed along four risk-adjusted performance metrics: inpatient length of stay ('efficiency'), waiting times ('access to care'), 28-day readmission rates and improvements in patient-reported health status after surgery (both 'clinical quality'). Each of these metrics has been the focus of recent health policy in England (Department of Health 2008; Department of Health 2012; Propper et al. 2008). We estimate multivariate multilevel models to account for the clustering of patients in providers and exploit the correlation of provider achievements across dimensions (Zellner 1962; Hauck and Street 2006). Empirical Bayes estimates of the provider-specific posterior

means and variance-covariance matrices are used to classify hospitals into three categories: dominant, dominated, and non-comparable. We quantify the uncertainty surrounding this classification in the form of Bayesian probability statements.

The study is the first to apply dominance criteria to multidimensional performance assessment of healthcare providers and derive appropriate confidence statements. Besides this, we make three further contributions to the empirical literature on hospital performance. First, we provide evidence about the correlations, and thus the potential for trade-offs, between a number of objectives that healthcare providers typically face. Previous research has focused predominantly on the association between hospital costs and mortality (see Hussey et al. (2013) for a review), largely ignoring other important dimensions such as waiting times or health-related quality of life. Second, in contrast to previous studies conducted at hospital level (e.g. Martin and Smith 2005), we focus on a single homogeneous patient population, thereby reducing the risk of ecological fallacy. Third, by exploiting novel data on pre-operative health status in addition to the co-morbidity markers that are usually available in administrative records, we are better able to isolate from case-mix differences the true impact that providers have on performance measures ('value added').

The remainder of this paper is structured as follows: In section 2 we set out the assessment framework in conceptual terms. Section 3 presents the empirical methodology and section 4 describes our data. We report results in section 5 and offer concluding comments in section 6.

2. Multivariate performance assessment using dominance criteria

Assume that a regulator, acting on behalf of stakeholders, seeks to determine the overall performance of a number of hospital providers. Let there be $k = 1, \ldots, K$ performance dimensions with observed achievement Y_k. Each achievement is determined by two factors, namely factors under the control of the provider θ_k and external production constraints X_k, so that

$$Y_k = f(X_k, \theta_k) \tag{1}$$

for each provider.

The parameter θ_k can be interpreted as the provider's contribution to achievement k over and above the circumstances in which they operate. This parameter is generally not directly observable and thus forms the target for inferences about performance. In order to isolate θ_k from X_k, the regulator must establish the contribution of production constraints to observed achievement by means of comparison with other providers, i.e. through risk-adjustment as applied in yardstick competition (Shleifer 1985).

Stakeholders derive utility from the providers' performance on each dimension, so that $U = U(\theta_1, \ldots, \theta_K)$, which is assumed to be monotonic in θ_k over the range of realistic values for all $k \in K$. The regulator has only limited knowledge about the characteristics of this utility function. This may be because there are multiple stakeholders with heterogeneous and/or unknown preferences. More specifically, the regulator has no information about the marginal utility $\partial U/\partial\theta_k$ that each stakeholder derives from achievements on each performance dimension, and hence the marginal rate of substitution (MRS) at which each stakeholder is willing to trade off performance on one dimension against that on another, i.e. $\partial\theta_k/\partial\theta_{k'}$ for $k \neq k^{'}$. However, the regulator has knowledge about the sign of $\partial U/\partial\theta_k$, i.e. whether achievements are expressed positively or negatively. To simplify the exposition, we assume from now on that achievements can be expressed so that utility increases in θ_k.

If only one performance dimension is assessed $(K = 1)$ or the MRS across multiple dimensions are known then achievements can be expressed as unidimensional (composite) scores. The regulator can then conduct either a *relative* or *absolute* assessment of performance. The first involves ranking the providers $j \in J$ according to their adjusted (composite) achievement θ_j, where $\theta_j > \theta_{j'}$ implies $U(\theta_j) > U(\theta_{j'})$ for $j \neq j^{'}$. This will result in a complete and transitive ordering of providers, assuming no ties. One can then designate a specific number of providers as performing well or poorly based on their relative ranking, e.g. whether they fall within a given percentile of the distribution. Goldstein and Spiegelhalter (1996) provide a discussion of the statistical challenges associated with this approach. Alternatively, providers' performances can be classified based on $\theta_j - \theta^*$ being larger or smaller than zero, where θ^* denotes an absolute performance standard to which providers are compared.[3] The latter is often employed in the context of quality performance assessment, e.g. with respect to standardised mortality after surgery (Spiegelhalter 2005; National Clinical Audit Advisory Group 2011).

When multiple performance dimensions are assessed $(K \geq 2)$ and the MRS are unknown, a complete and transitive ordering of providers is no longer guaranteed and relative assessments are unfeasible. As a result, it becomes impossible to identify providers that perform well or poorly in terms of stakeholders'

[3] Note that, when no external standards are specified, performance standards are typically based on the performance of all organisations, i.e. an internal performance standard (Shleifer 1985; National Clinical Audit Advisory Group 2011). Hence, a provider will be considered to perform well when the observed achievement is better than a reference value derived from all providers. In many cases, this reference value is simply the average across all providers, i.e. $\theta^* = \frac{1}{J}\sum\theta_j$.

aggregate utility. This is a well-known problem in the field of welfare economics and consumer theory (Boadway and Bruce 1984; McGuire 2001). However, some combinations of performance levels may be strictly preferable (dominant) or inferior (dominated) to other combinations, leading to a partial ordering of provider. As an analogue to the Pareto dominance criteria we can formalise the following general dominance classification rules[4]:

A provider either

1. *dominates* the comparator if $\theta_{jk} \geq \theta_{j'k}$ for all $k \in K$ and $\theta_{jk} > \theta_{j'k}$ for some $k \in K$, or
2. *is dominated* by the comparator if $\theta_{jk} \leq \theta_{j'k}$ for all $k \in K$ and $\theta_{jk} < \theta_{j'k}$ for some $k \in K$, or
3. is *non-comparable* to the comparator if $\theta_{jk} \geq \theta_{j'k}$ for some $k \in K$ and $\theta_{jk} \leq \theta_{j'k}$ for the remaining $k \in K$,

where $j \neq j^{'}$ and $\theta_{j'k}$ denotes the performance level of the comparator, which may be either another provider or an absolute internal or external performance standard θ^{*}.

[4] Devlin et al. (2010) propose the use of a similar classification system to compare EQ-5D health profiles over time without resorting to making strong assumptions about patients' preferences.

3. Methodology

3.1. Empirical approach

The aims of the empirical analysis are to obtain estimates of providers' performance θ_{jk} and of the correlation of θ_{jk} across each of the $K = 1, \ldots, 4$ performance dimensions, and to classify providers according to the dominance classification set out in section 2. We estimate multivariate multilevel models (MVMLMs) with achievement score Y_{ijk} observed for patients $i = 1, \ldots, n_j$ who are clustered in hospitals $j = 1, \ldots, J$. Multilevel (i.e. random intercept) models have become a staple tool in the field of performance assessment and allow us to i) adjust achievements for differences in case-mix across providers, ii) decompose unexplained variation in achievement into random (within-provider) variation at patient level and systematic (between-provider) variation at provider level, and iii) obtain more reliable (precision-weighted or shrunken) estimates of performance (Goldstein 1997; Normand et al. 1997; Ash et al. 2012).

The multivariate nature of the data is taken into account through correlated random terms at each level of the hierarchy. These random terms are assumed to be drawn from multivariate normal distributions (MVN) with unconstrained variance-covariance matrices (Zellner 1962; Hauck and Street 2006). Allowing for correlation across achievements is beneficial for several reasons. First, we can construct multivariate hypothesis tests of parameters of interest that take into account the correlation between dimensions and achieve correct coverage probabilities. We discuss this in detail below. Second, we can achieve efficiency gains and obtain more precise estimates of relevant parameters if either the components of X_{ijk} differ across k or non-identity link functions are employed for at least some of the regression equations (Zellner 1962; Thum 1997; Bailey and Hewson 2004). Finally, by utilising a maximum likelihood estimator, data about achievements that are missing for any particular performance domain can be assumed missing at random conditional on all modelled covariates *and* achievements (Little and Rubin 1987; Goldstein 1986).

Hospital achievements are measured using two continuous and two binary variables. In order to ascertain the conditional normality of error terms as imposed by the MVN assumption[5], we apply appropriate transformations (e.g. logarithmic) for the continuous achievement variables and specify probit models for the binary achievement variables. The latter can be motivated by considering each binary achievement variable as the observed realisation of a latent truncated Gaussian variable.

The empirical model to be estimated is specified as

$$Y^*_{ijk} = \alpha_k + X'_{ijk}\beta_k + \theta_{jk} + \epsilon_{ijk} \tag{2}$$

with $Y^*_{ijk} = f(Y_{ijk})$ for $k = 1, 2$ and

$$Y_{ijk} = \begin{cases} 1 \text{ if } Y^*_{ijk} > 0 \\ 0 \text{ if } Y^*_{ijk} \leq 0 \end{cases}$$

for $k = 3, 4$.

The variable Y_{ijk} denotes the observed outcome, Y^*_{ijk} is the corresponding latent underlying variable, $f(.)$ is a transformation function chosen to normalise the conditional distribution of ϵ_{ijk}, X_{ijk} is a

[5] In principle it is possible to use other multivariate distributions such as multivariate gamma. However, such models are not typically implemented in standard statistical software packages and are therefore rarely used in practice.

vector of explanatory variables whose components may differ across dimensions, α_k is an intercept term, θ_{jk} denotes a random effect at provider level and ϵ_{ijk} denotes the random error term at patient level. Both random terms are assumed to be MVN distributed with mean vector zero and a $K \times K$ variance-covariance matrix, so that $\theta_{jk} \sim MVN(0, \Sigma)$ with

$$\begin{aligned} E(\theta_{jk}) &= 0 \\ var(\theta_{jk}) &= \tau_k^2 \\ cov(\theta_{jk}, \theta_{jk'}) &= \rho_\theta \tau_k \tau_{k'} \end{aligned}$$

for all $k \neq k^{'}$, and similarly $\epsilon_{ijk} \sim MVN(0, \Omega)$ with

$$\begin{aligned} E(\epsilon_{ijk}) &= 0 \\ var(\epsilon_{ijk}) &= \sigma_k^2 \text{ for } k = 1, 2 \\ var(\epsilon_{ijk}) &= 1 \text{ for } k = 3, 4 \\ cov(\epsilon_{ijk}, \epsilon_{ijk'}) &= \rho_\epsilon \sigma_k \sigma_{k'} \end{aligned}$$

for all $k \neq k^{'}$. The model reduces to a set of univariate models if all off-diagonal elements of Σ and Ω are zero, i.e. achievements are uncorrelated conditional on observed patient factors.

Estimation was performed in MLwiN 2.32 called from within Stata 13 using the `runmlwin` programme (Leckie and Charlton 2013).

3.2. Classification of provider effects and multivariate hypothesis tests

We compare providers against a common absolute performance standard, here defined as the expected performance of a (hypothetical) hospital of average performance α_k, i.e. the conditional mean. We base our assessment of provider performance on estimates of θ_{jk}, which represent the provider-specific deviation from this benchmark. These parameters are not directly estimated in a random effects framework but can be recovered in post-estimation using Empirical Bayes predictions techniques (Skrondal and Rabe-Hesketh 2009). We stack performance estimates into vector coordinates to denote the provider's location in the k-dimensional performance space with the origin being normalised to zero. A provider's dominance classification is then determined by comparing its estimated adjusted achievements to that of the performance standard across all four dimensions simultaneously. This leads to three possible classifications: dominant, dominated, or non-comparable.

In order to quantify the uncertainty around these possible classifications we take a Bayesian perspective and calculate the posterior probability that a given provider truly dominates [is dominated by; non-comparable to] the multidimensional performance standard. This involves calculating the area under the MVN probability density function that covers each of the three possibilities, for each provider.[6] Figure 1a illustrates this for the two-dimensional case with two highly correlated bivariate normal distributed achievements ($\rho = 0.6$). The centroid of the density is given by X and the ellipse shows the central 95% of this density. The density is dissected by two lines which intersect at the benchmark. The density covered by the areas A and B equal the probability of *dominating* or *being dominated by* the benchmark, whereas the density covered by area C gives the probability for the *non-comparable* outcome. To calculate

[6] Our problem is similar to that encountered in the context of cost-effectiveness analysis, where one wishes to compute the probability that a new treatment is cost-effective for a given level of willingness to pay (Van Hout et al. 1994; Briggs and Fenn 1998; O'Hagan et al. 2000).

these probabilities, we follow the simulation approach of O'Hagan et al. (2000). Our simulation involves drawing S repeated samples from the MVN posterior distribution of the provider-specific Empirical Bayes estimates of the mean vector θ_j and associated variance-covariance matrix Σ_j. We then apply the dominance criteria to each simulation and calculate posterior probabilities by averaging across simulations. Formally,

$$Pr(\text{dominant } |J = j) = \frac{1}{S} \sum_{s=1}^{S} \prod_{k=1}^{4} I(\theta_{jk}^{s} > 0) \tag{3}$$

$$Pr(\text{dominated } |J = j) = \frac{1}{S} \sum_{s=1}^{S} \prod_{k=1}^{4} I(\theta_{jk}^{s} < 0) \tag{4}$$

and by construction

$$Pr(\text{non-comparable } |J = j) = 1 - (Pr(\text{dominant } |J = j) + Pr(\text{dominated } |J = j)) \tag{5}$$

where S is the total number of simulations (here $S = 10,000$), θ_{jk}^{s} denotes the simulated provider-effect in simulation s, and I is an indicator function that takes the value of one if the condition is true and zero otherwise. This approach has several advantages over a series of univariate assessments: Most importantly, it accounts for the correlation between performance dimensions and thus achieves correct coverage of the confidence region (Briggs and Fenn 1998). Figure 1b illustrates the difference between probability statements if performances on both dimensions are incorrectly assumed to be independent. The dashed line outlines the resulting 'confidence box', which is formed by the end points of two independent 95% confidence intervals that are adjusted for multiple testing. Furthermore, because we make probability statements about a single quantity of interest, the provider's location in the k-dimensional performance space, we avoid such issues of multiple testing.

3.3. Risk-adjustment

Perhaps the primary reason that observed achievements differ across hospitals is because they treat different types of patients. To account for this, we develop specific risk-adjustment models for three of the performance dimensions. Based on previous research (Gutacker et al. 2013; Street et al. 2014), we identify a set of 'core' variables common to all models: patient age, gender, pre-treatment health status, primary diagnosis (coded as osteoarthritis (ICD-10: M15-19), rheumatoid arthritis (ICD-10: M05-06), or other), comorbidity burden, socio-economic status, and year of treatment. Other variables considered were time with symptoms, whether the patient lived alone, whether the patient required assistance filling in the PROM questionnaire, or whether she considered herself disabled.[7] Finally, in the length of stay model, we controlled for the healthcare resource group (HRG, the English equivalent of Diagnosis Related Groups) to which the patient was allocated.

Preliminary modelling of potential risk-adjusters was conducted on the basis of univariate multilevel regression models and visual inspection of LOWESS plots (for continuous variables) and box plots (for categorical variables). A significance level of $p < 0.05$ was required for variables to be retained. All continuous variables were first added linearly to the regression model and we subsequently explored whether squared terms improved the fit of the model. As expected, our exploratory work confirmed the importance of all core variables in explaining variation in each of the three performance dimensions.

[7] We only consider information contained in the pre-operative questionnaire since the e.g. need for assistance in filling in the post-operative questionnaire may be endogenous to the outcome of the care process.

(a) Areas covered if acknowledging correlation

(b) Difference between areas if achievements are assumed independent ('confidence box')

Legend: X denotes the centroid of the density. The solid ellipsoid line shows the inner 95% of the bivariate density with $\rho = 0.6$, whereas the dashed line denotes the density covered by the confidence box that is formed by two independent 95% confidence intervals. The horizontal and vertical axes intersect at the benchmark and dissect each density into four areas, where the covered density of the area reflects the probability of dominating the benchmark (A), being dominated by the benchmark (B) or being non-comparable to the benchmark (C) (left panel).

Figure 1: Example of area of probability density plane covered under different assumptions about the dependence of achievement scores

Time with symptoms, assistance and living alone did not explain variation in the probability of being re-admitted and were thus not included in the final model. Non-linear effects were found for age (all performance dimensions) and pre-treatment health status (only length of stay and post-operative OHS).

No risk-adjustment was performed in the analysis of waiting times because providers are expected to manage their waiting lists so as to balance high priority cases and those with less urgent need for admission.

3.4. Endogeneity due to patient selection of healthcare provider

Patients in the English NHS have a right to choose their provider of inpatient care for most elective procedures. This may lead to bias in the estimates of hospital performance if both the choice of hospital and the achievements for an individual patient are driven by common underlying factors that are not controlled for as part of X_{ijk}. This may arise if patients self-select into hospitals based on unobserved characteristics or providers cream-skim (Gowrisankaran and Town 1999; Geweke et al. 2003). Examples include unobserved severity, health literacy or other factors that enter the personal health production function and are also determinants of hospital choice.

In order to test for bias due to patient selection and to obtain correct estimates of hospital performance, we estimate the model in (2) and perform two-stage residual inclusion (2SRI) as suggested by Terza et al. (2008). In the first stage, we estimate a multinomial choice model of hospital choice, where choice is assumed to be determined by the straight-line distance[8] from the patient's residence to the provider, an unobserved patient effect and random noise. Distance is commonly chosen in the literature as an instrumental variable as it is a major driver of hospital choice and is exogenously determined, on the reasonable assumption that patients do not choose where to live based on hospital performance (Gowrisankaran and Town 1999). The residual from this regression captures both the unobserved patient effect and random noise. In the second stage, we enter this residual as an additional regressor into each of the four achievement regression models. If the coefficients on the first-stage residuals are estimated to be statistically significantly different from zero this provides evidence of selection bias and the need for adjustments based on 2SRI (Terza et al. 2008).

[8] We also include distance^2 and distance^3 as well as an indicator for whether the hospital is the closest alternative. Hospitals with less than 30 patients were removed from the choice set. The patient's residence was approximated by the centroid of the lower super output area (LSOA) in which the patient lives. LSOAs are designed to include approximately 1,500 inhabitants, i.e. they are substantially smaller than US ZIP codes.

4. Data

Our primary source of data is the Hospital Episode Statistics (HES) data warehouse, which contains detailed inpatient records for all patients receiving NHS-funded care in England. We extract information on all patients undergoing unilateral hip replacement (identified through the primary procedure code; see Department of Health (2008)) in the period April 2009 to March 2012.[9] Patients were excluded if they were aged 17 or younger at the time of admission, underwent revision surgery, were admitted as emergencies or day-cases, or if information on important risk-adjustment variables was missing. Patients were also excluded if they attended a provider that treated fewer than 30 patients in the same financial year. We record any hospital admission occurring within 28 days after the initial admission for hip replacement surgery. All linkage was achieved using unique patient identifiers.

For each patient, we extract information on demographics and socio-economic background, medical characteristics and information pertaining to the admission process and the hospital stay itself. These data are used to construct three achievement measures: i) inpatient length of stay (top-coded at the 99th percentile), ii) emergency re-admission within 28 days of discharge for any condition (coded as 0=not re-admitted, 1=re-admitted), and iii) waiting time, measured as the time elapsed between the surgeon's decision to admit and the actual admission to hospital. Waiting time is categorised into waits of no more than 18 weeks (=0) and waits exceeding 18 weeks (=1) to mirror the contemporaneous waiting time performance standard in the English NHS.[10] We also derive the following risk-adjustment variables from HES: age, sex, comorbidity burden as measured by individual Elixhauser comorbidity conditions recorded in secondary diagnosis fields (Elixhauser et al. 1998), number of emergency admissions to hospital within the last year (coded as 0=none, 1=one or more), and patients' approximate socio-economic status based on level of income deprivation in the patient's neighbourhood of residence as measured by the Index of Multiple Deprivation 2004 (Noble et al. 2006).

We link HES records to data from the national Patient Reported Outcome Measures (PROM) survey. This survey invites all patients undergoing unilateral hip replacement to report their health status before and six months after surgery using the Oxford Hip Score (OHS) (Dawson et al. 1996).[11] The OHS is a reliable and validated measure of health status for hip replacement patients and consists of twelve questions regarding functioning and pain. For each item, the patient is asked to respond on a five-item scale. These items are summed up to generate an index score ranging from 0 (worst) to 48 (best). The post-operative OHS forms the fourth achievement measure and the pre-operative OHS score is used to control for initial health status at admission. Because we observe pre-operative health status in addition to the co-morbidity markers that are usually available in administrative records, our estimates of performance are more likely to indicate the true impact that providers have on performance measures ('value added') rather than reflect residual case-mix differences. The PROM survey also gathered additional information on duration of problems, and whether the patient lives alone, considered herself disabled, or required help filling in the questionnaire. Pre-operative survey responses are collected by paper questionnaire

[9] HES records activity at the level of 'finished consultant episodes' (FCEs) and we link consecutive episodes within the hospital stay and across hospital transfers to form continuous inpatient spells (CIPS). A CIPS is deemed complete when the patient is discharged from one provider and not re-admitted to another provider within 2 days.

[10] The current performance standard is defined in terms of proportion of patients exceeding a waiting time of 18 weeks between the GPs referral and the admission. Unfortunately, data on the time elapsed between the GPs referral and the surgeon's decision to admit are not recorded in HES. Our performance estimates will therefore be overstated.

[11] All patients are also invited to fill in the EuroQol-5D (EQ-5D) questionnaire, a generic health-related quality of life instrument (Brooks 1996). However, we focus on the OHS as it is better approximated by a continuous distribution and we do not seek to make comparisons across disease areas. Furthermore, the OHS is the relevant outcome measure for the newly introduced best practice tariff (a pay-for-performance scheme) in the English NHS that was introduced in April 2014 (Monitor and NHS England 2013). Previous comparisons have demonstrated that performance assessments based on the EQ-5D and OHS lead to similar conclusions (Neuburger et al. 2013).

during the last outpatient appointment or on the day of admission, whereas follow-up responses are collected via mailed survey to the patient's home address. Participation in the PROM survey is voluntary for patients but mandatory for all providers of NHS-funded care. Approximately 60% of patients returned completed pre-operative questionnaires that can be linked to HES (Gutacker et al. 2015).

5. Results

5.1. Descriptive statistics

The estimation sample consists of 95,955 patients treated in 252 providers during April 2009 and March 2012. Table 1 presents descriptive statistics. Patients are on average 67 years old, and approximately 41% of patients are male. The majority (68%) report having had problems with their hip joint for 1 to 5 years, although about 8% of patients experienced symptoms for more than 10 years and 14% reported problems for less than 1 year. Approximately 39% of patients classify themselves as having a disability, and 27% live alone.

Table 1: Descriptive statistics

Description	N	Mean	SD
Achievement measures (Dependent variables)			
Post-operative OHS	81,336	38.50	9.21
Length of stay (in days)	95,878	5.36	3.75
Waiting time $>$ 18 weeks (1=yes, 0=no)	92,154	0.17	0.38
28-day emergency readmission (1=yes, 0=no)	95,955	0.05	0.22
Patient characteristics (Control variables)			
Patient age (in years)	95,955	67.43	11.29
Patient gender (1=male, 0=female)	95,955	0.41	0.49
Pre-operative OHS	95,955	17.66	8.28
Primary diagnosis			
Osteoarthritis (1=yes, 0=no)	95,955	0.93	0.25
Rheumatoid arthritis (1=yes, 0=no)	95,955	0.01	0.07
Other (1=yes, 0=no)	95,955	0.06	0.24
Number of Elixhauser comorbidities			
0	95,955	0.35	0.48
1	95,955	0.29	0.45
2-3	95,955	0.26	0.44
4+	95,955	0.10	0.31
Previously admitted as an emergency (1=yes, 0=no)	95,955	0.08	0.28
Socio-economic status	95,955	0.12	0.09
Disability (1=yes, 0=no)	95,955	0.39	0.49
Living alone (1=yes, 0=no)	95,955	0.27	0.44
Assistance (1=yes, 0=no)	95,955	0.21	0.41
Symptom duration			
$<$ 1 year	95,955	0.14	0.35
1 - 5 years	95,955	0.68	0.47
6 - 10 years	95,955	0.11	0.31
$>$ 10 years	95,955	0.08	0.26
Healthcare Resource Group			
HB12C - category 2 without CC	95,955	0.77	0.42
HB11C - category 1 without CC	95,955	0.10	0.29
HB12B - category 2 with CC	95,955	0.07	0.26
HB12A - category 2 with major CC	95,955	0.04	0.19
HB11B - category 1 with CC	95,955	0.01	0.11
other	95,955	0.02	0.12

Legend: N = Number of observations, SD = Standard deviation; OHS = Oxford Hip Score; CC = complications or co-morbidities.
Notes: Healthcare Resource Groups refer to major hip procedures for non-trauma patients in category 1 (HB12x) or category 2 (HB11x). Socio-economic status is approximated by the % of neighbourhood residents claiming income benefits. This characteristics is measured at neighbourhood level (lower super output area (LSOA)).

Figure 2 illustrates the empirical distributions of the achievement variables on their untransformed scales. The average post-operative OHS is 38.5 (SD=9.2) and the average length of stay is 5.4 days (SD=3.8), with both distributions showing substantial skew. Approximately 5.2% of patients were readmitted to hospital within 28 days of discharge, and about 17.5% of patients waited longer than 18 weeks to be admitted to hospital. There is a substantial proportion of missing responses in terms of post-operative OHS (15.2%) and, to lesser degrees, waiting time (4.0%) and length of stay (0.1%). Conversely, emergency re-admission status is recorded for all patients.

5.2. Provider heterogeneity and correlation between performance dimensions

From the estimated variance-covariance matrices Σ and Ω we can calculate the correlation across performance estimates.[12] The lower off-diagonal in Table 2 shows the correlation between performance estimates at provider level, whereas the upper off-diagonal shows the correlation at patient level. Bold numbers indicate that the correlation coefficient is statistically significantly different from zero ($p<0.05$; Huber-White standard errors).

Table 2: Correlation between performance dimensions

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.34**	1.00	***-0.02***	***-0.07***
Waiting time $>$ 18 wks (3)	**0.26**	**-0.31**	1.00	*0.00*
28-day emergency readmission (4)	0.03	**-0.49**	0.16	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

We focus our discussion on the correlation between performance dimensions at provider level. Our results suggest significant correlations for four combinations of dimensions. Hospitals with shorter length of stay also realise better post-operative health status for their patients (ρ = -0.34; SE = 0.067; $p<0.001$). This is consistent with findings from randomised controlled trials that tested the effectiveness of so-called 'fast track' or 'enhanced recovery' pathways and found that hospitals that mobilise patients sooner after surgery were able to discharge them quicker and achieve better post-operative outcomes (Husted et al. 2008; Larsen et al. 2008; Paton et al. 2014). We also find evidence to suggest that hospitals with shorter length of stay also have a lower proportion of patients waiting more than 18 weeks to be admitted (ρ = 0.26; SE = 0.065; $p<0.001$), suggesting better management of capacity and of their waiting lists. This would be consistent with a queuing model of limited bed capacity, where prospective patients cannot be admitted until current patients are discharged. Hospitals that have better post-operative health outcomes also tend to have a lower proportion of patients waiting for more than 18 weeks (ρ= -0.31; SE = 0.071; $p<0.001$). Finally, the correlation between post-operative health status and probability of an emergency readmission within 28 days is negative and statistically significant (ρ = -0.49; SE = 0.078; $p<0.001$). Overall, these correlations indicate that inferences based on a series of univariate assessments would likely be misleading and that our MVMLM is preferable for this empirical analysis of provider performance.

[12] All achievements are adjusted for case-mix. The estimated coefficients on risk-adjustment variables and associated standard errors are not the focus of this study and are reported in Table A1 in the Appendix. The first-stage residuals from the selection equation are jointly statistically significant ($\chi^2(4)$ = 14.97; $p<0.01$) when entered into the main equations, suggesting that self-selection into hospital may bias performance estimates if uncontrolled for (see Table A2 in the Appendix for first-stage estimates).

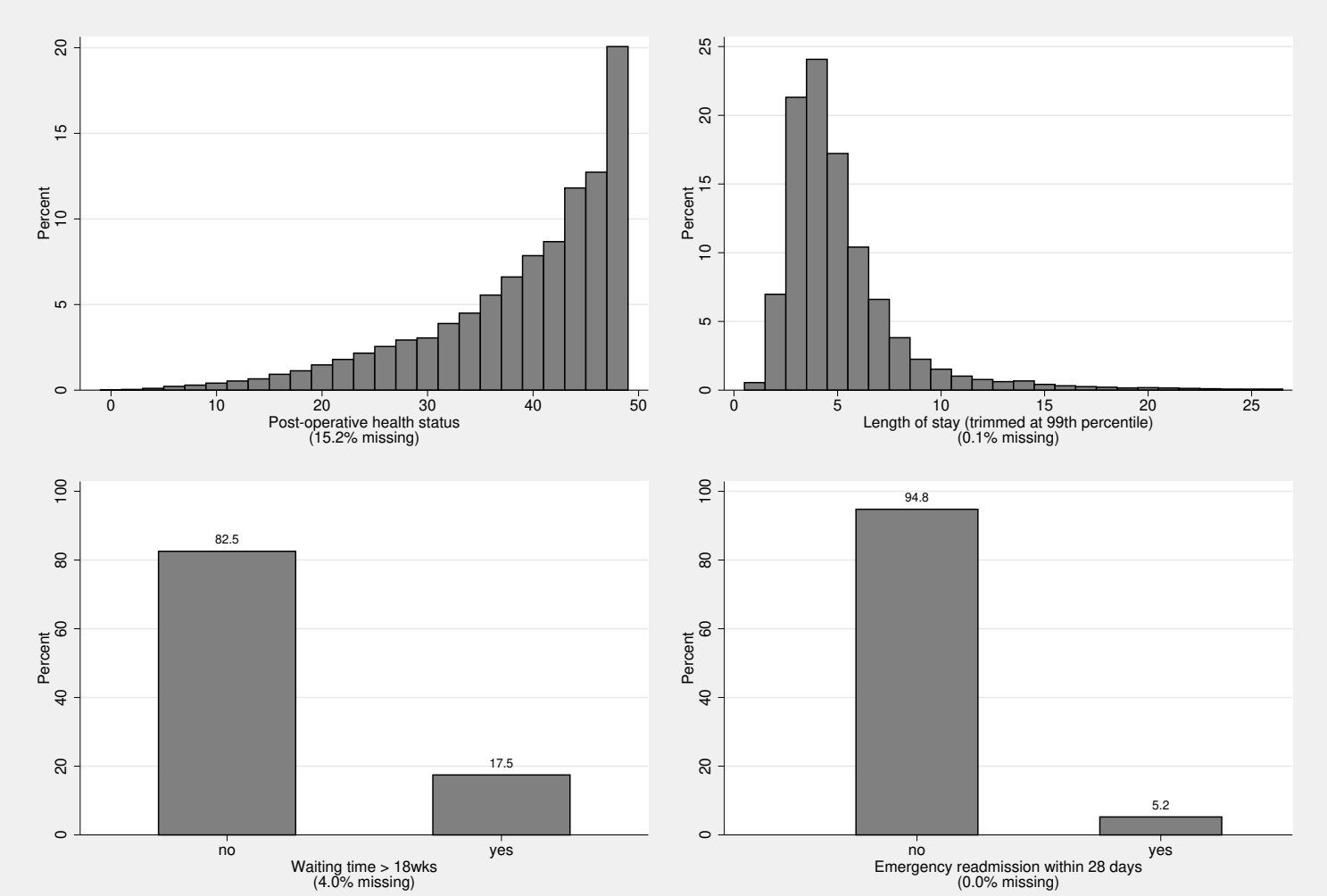


Figure 2: Empirical distribution of unadjusted achievement scores

It is also of interest to understand how much of the observed variability in adjusted achievement scores can be attributed to providers (Hauck et al. 2003). We calculate the intraclass correlation coefficient (ICC)[13] for each of the four performance dimensions with confidence intervals formed by the delta method. The largest ICC is observed for waiting times with 27.4% (SE = 0.020; p<0.001) of unexplained variation in achievements occurring between providers, followed by length of stay with approximately 13.3% (SE = 0.011; p<0.001). In contrast, the ICCs on the achievements post-operative OHS (1.7%; SE = 0.002; p<0.001) and emergency readmission (2.2%; SE = 0.003; p<0.001) are substantially smaller; implying that providers have less influence over these performance dimensions.

We have conducted sensitivity analyses with respect to a number of modelling choices (results are reported in Appendix Tables A3 to A5): First, we excluded privately owned and operated providers (so called 'independent sector treatment centres' (ISTCs)) as these may be argued to operate under different production constraints (see below). The estimated covariance terms in Σ are somewhat attenuated and the correlations of waiting time with length of stay (p=0.174) and post-operative health status (p=0.857) are no longer statistically significant. Second, we included additional regressors based on patient risk factors averaged at provider level to correct for potential bias arising from correlation between X_{ij} and the hospital random effects (Mundlak 1978).[14] Due to convergence problems, we restricted this to patient age, pre-operative PROM score and level of income deprivation. Again, covariance terms are smaller in size but remain statistically significant. Finally, we restricted the risk-adjustment to variables that can be derived from routine administrative data, i.e. we excluded all variables based on the PROM survey. Results are robust to this omission.

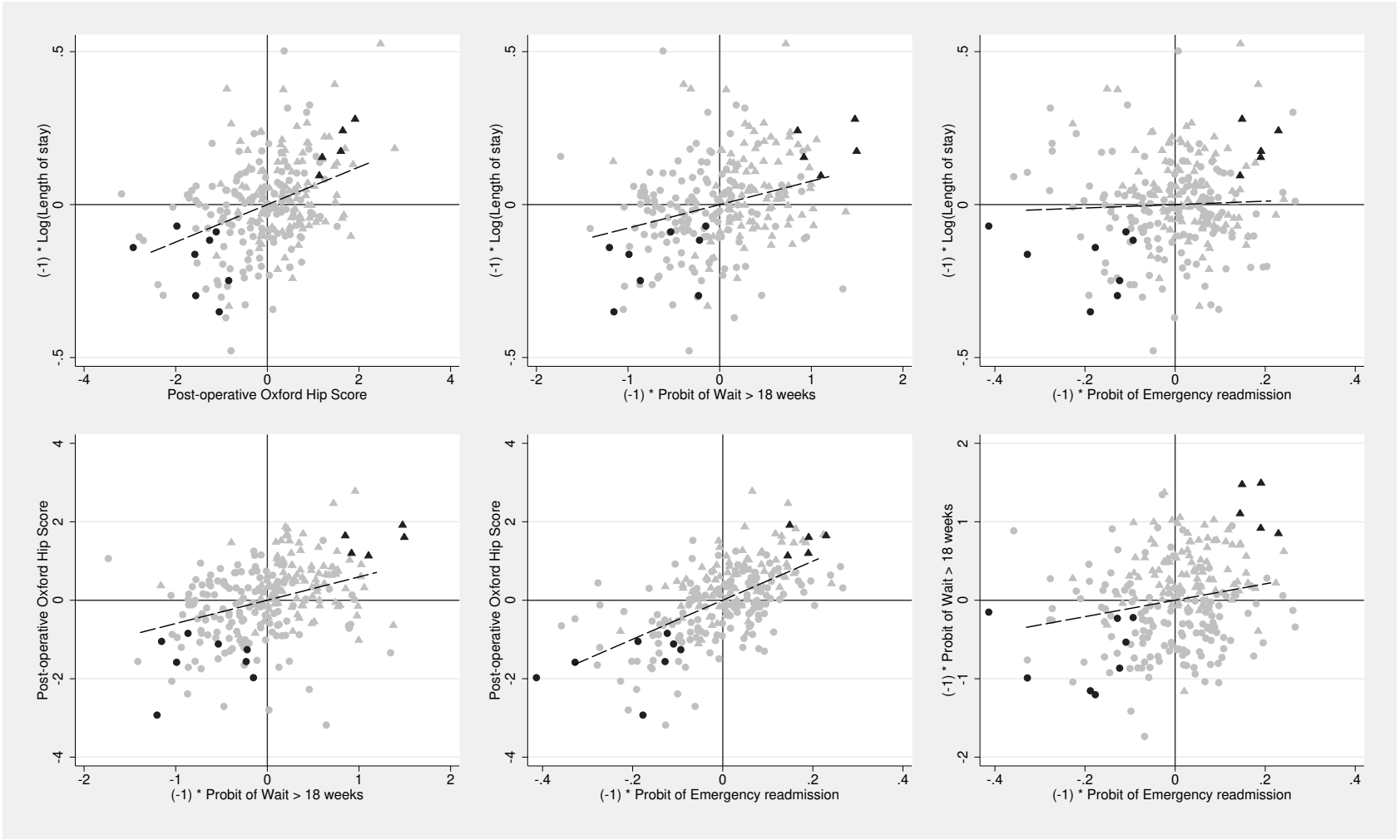
5.3. Provider performance assessment

We now turn to the assessment of multidimensional provider performance. Figure 3 shows the location of each provider in the four-dimensional performance space, where each panel presents scatter plots for two dimensions. The axes for all performance dimensions except post-operative health status are reversed (i.e. multiplied by -1) so that higher scores indicate better performance. Hence, providers in the NE quadrant perform better than the benchmark on both dimensions, whereas those in the SW quadrant perform worse. Providers that dominate or are dominated by the multidimensional benchmark with at least 90% probability are highlighted as darker points.

Figure 3 shows that we identify five dominant and eight dominated providers at a probability level of 90%. It turns out that all dominant providers are privately owned and operated treatment centres that perform mainly orthopaedic procedures, here marked as triangles, whereas all dominated providers are public NHS providers, marked as circles, that provide a wider mix of services, including emergency care. Note however that not all ISTCs are located in the NE quadrant, and not all NHS providers are located in the SW quadrant. To test whether the observed performance advantage of ISTCs also holds on average, we re-estimated the models and included an indicator variable for private ownership. We found statistically significant effects on length of stay (beta=-0.100; SE = 0.020; p<0.001), post-operative health status (beta=1.205; SE = 0.157; p<0.001), probability of being readmitted (beta=-0.084; SE = 0.072; p<0.001), and the probability of waiting longer than 18 weeks (beta=-0.820; SE = 0.030; p=0.007). Ideally one

[13] The ICC for performance dimension k is $ICC_k = \frac{\tau_k^2}{\tau_k^2+\sigma_k^2}$.

[14] This bias is likely to be small. We compared coefficient estimates from fixed and random effects estimators using Hausman tests and found little practical difference between those estimates, although the tests all rejected the assumption of unbiasedness for the random effects approach. This is likely to be due to our large sample, where within effects swamp between effects and the Hausman test is over-powered. Results are available on request.



Notes: Each of the six panels shows bivariate plots of performance estimates. Higher scores imply better performance. Triangles indicate privately operated providers and circles indicate NHS providers.

Figure 3: Multidimensional performance estimates

Table 3: Number of dominant/dominated providers under different estimation approaches and assumptions about the correlation between performance dimensions

Probability threshold Pr^*	(1) Univariate		(2) Intermediate multivariate		(3) Full multivariate	
	Dominant	Dominated	Dominant	Dominated	Dominant	Dominated
0.50	5	8	7	10	24	30
0.80	2	3	5	5	12	18
0.90	1	1	2	2	5	8
0.99	0	0	0	1	1	1

(1) Univariate approach - separate univariate models are estimated for each of the four performance dimensions and providers are considered dominant [dominated] if the independent probability of being dominant [dominated] exceeds $1-(1-Pr^*)/4$ on *each* of the four dimensions.
(2) Intermediate multivariate approach - multivariate model is estimated and providers are considered dominant [dominated] if the independent probability of being dominant [dominated] exceeds $1-(1-Pr^*)/4$ on *each* of the four dimensions. Correlation between performance dimensions is exploited in the estimation stage but ignored when forming probability statements.
(3) Fully multivariate approach - multivariate model is estimated and providers are considered dominant [dominated] if the probability of being dominant [dominated] on all four dimensions jointly exceeds Pr^*. See section 3.2 for details.

would compare dominant ISTCs and dominated NHS hospitals across a range of characteristics (e.g. staffing ratios, experience of surgical teams, profit margin, etc.) to generate hypotheses about the likely causal factors underlying those performance differences. Unfortunately, data limitations, especially with respect to ISTCs, prevent us from doing so.

5.4. Comparison with approaches based on series of univariate probabilities

It is instructive to compare the results from our MVMLM assessment with two alternative approaches: 1) a series of four univariate multilevel regressions, and 2) an 'intermediate' MVMLM regression that takes into account the correlation between achievements during the estimation stage but treats performance estimates as independent. In both cases a provider is judged to be dominant [dominated] if all four individual probabilities of exceeding [falling short of] the benchmarks are greater or equal to a specified probability threshold ('confidence box approach'). The second approach can thus be seen as an intermediate between a simple univariate approach and the full multivariate approach employed in this study.

The univariate and intermediate multivariate approach both involve comparing four independent probabilities against a threshold value, which would lead to inflated risk of classifying providers as dominant [or dominated] when they are not (type I error). We adopt here the Bonferroni correction to adjust for multiple comparisons, i.e. we require (1-(1-Pr^*)/4)*100% probability on each of the four dimensions to designate a provider as dominant/dominated, where Pr^* equals the desired level of certainty.

Table 3 shows the number of provider identified as dominant/dominated under each of these approaches. At a probability threshold of 90% (Pr^*=0.9), the univariate and intermediate multivariate both identify just one or two dominant and dominated providers, which is fewer than the full MVMLM. The intermediate multivariate approach is more efficient than the univariate approach. This becomes apparent when applying an 80% threshold. At this probability threshold the univariate assessments identify two dominant and three dominated providers, whereas the intermediate MVMLM identifies five dominant and five dominated providers. The full MVMLM approach identifies 12 dominant and 18 dominated providers at the 80% threshold.

6. Discussion

Rarely are stakeholders explicit about the valuations they attach to different dimensions of performance, nor are these valuations likely to be identical. This renders the construction of a composite performance indicator that is appropriate for all audiences unfeasible. To circumvent this, we have set out a methodology for comparing healthcare providers in terms of their performance across a range of dimensions in a way that does not require valuation of each dimension and is consistent with economic theory. Building on previous literature, we analyse relative provider performance for each dimension and allow for correlation across dimensions (e.g. Hauck and Street 2006; Martin and Smith 2005; Hall and Hamilton 2004). We extend this literature by employing dominance criteria to compare providers against a multidimensional benchmark, and by constructing multivariate (rather than univariate) hypothesis tests of parameters that account for correlation between dimensions and thereby achieve correct coverage probabilities. Failure to perform multivariate tests can lead to incorrect inferences about multidimensional performance as we illustrate.

We have applied our MVMLM approach to study the performance of English providers of care to patients having hip replacement. By focusing on a single procedure, we can draw more robust conclusions about performance than studies conducted at hospital level. Our use of patient-level data allows us to employ multilevel models to control for a diverse range of patient characteristics and, thereby, to isolate the provider's impact on observed achievements. We study four dimensions of performance, namely long waiting times ($>$18 weeks), length of stay, 28-day readmission rates, and patient-reported health status after surgery. Achievements on some of these dimensions are correlated, implying that our multivariate estimation framework is appropriate. Our results do not suggest trade-offs between achievements on the four performance dimensions we studied. Instead, we observe positive, albeit weak, correlations. We wish to stress that these results do not necessarily imply a causal relationship between achievements, although some of our findings confirm those of randomised controlled trials conducted in routine care settings.[15] Nevertheless, this suggests that pairs of achievements are either a) driven by common underlying factors that enter both production functions, such as organisational effort, or b) that achievements on one dimension enable achievements on another. This information is of interest itself as it informs the debate whether incentive schemes can be simplified to reward providers on a subset of correlated measures, as suggested by Glazer et al. (2008), or whether regulators should instead ascertain performance across all individual performance dimensions of interest.

Our estimation yields, for each provider, one performance estimate per performance dimension, which together form a provider's performance profile. To translate this profile into a single statement about performance we employ a set of dominance criteria and classify providers into three groups: (i) dominant providers, which are 'positive deviants' that exhibit outstanding performance across all performance dimensions; (ii) dominated providers, which are 'negative deviants' with sub-standard performance; and (iii) the remainder. In this study of patients having hip replacement, all dominant providers were found to be privately operated treatment centres specialising in elective (i.e. non-emergency) hip and knee replacement, while all dominated providers were public NHS providers providing a wide range of services. ISTCs have previously been found to achieve on average better health outcomes than public providers (Browne et al. 2008; Chard et al. 2011) and to discharge patients earlier (Siciliani et al. 2013), and we can confirm these findings in our data. This may be the result of a more stream-lined production process: ISTCs typically focus exclusively on elective orthopaedic procedures, such as hip and knee replacement, whereas NHS providers offer a wide range of service, including emergency care. If the organisational

[15] Importantly, these trials also provide evidence on the *direction* of the causal effect, i.e. what causes what.

set-up of ISTCs allows them to specialise almost exclusively on treating particular types of patient, this may result in performance advantages. Our data do not allow us to unpack the reasons for the observed performance further, and we stress that performance assessment results should form the starting point for further investigations involving site visits and qualitative analysis (Bradley et al. 2009). As with most regression analyses, general differences between types of providers can be identified using conditional mean comparisons, in which indicator variables are used to specify provider types. But our approach also allows us to identify positive and negative deviants *within* these broad categories. This is important as otherwise regulatory efforts may be accidently directed at those ISTCs that are found to perform relatively poorly.

The appeal of the dominance approach lies in the absence of strong assumptions about the various stakeholders' utility functions and its ability to reduce multiple performance estimates into a single assessment. However, this comes at a price. Because the approach requires providers to perform better than the benchmark on *all* dimensions, there is no scope to compensate for average or poor performance on one dimension through excellent performance on another. This very strict yardstick is difficult to achieve and so we identify only a small number of providers as dominant or dominated. Also, as the number of objectives under consideration increases it becomes increasingly more difficult to satisfy the dominance criteria (Pedraja-Chaparro et al. 1999). Nevertheless, although we have illustrated our methodology by analysing only four dimensions, it is generalisable to multiple dimensions.

These qualifications not withstanding, we advocate the dominance approach to multidimensional performance assessment as a useful addition to regulators' toolboxes.

References

Appleby, J., E. Poteliakhoff J, K. Shah and N. Devlin (2013). 'Using patient-reported outcome measures to estimate cost-effectiveness of hip replacements in English hospitals'. *Journal of the Royal Society of Medicine*, 106: 323–331.

Arrow, K. (1963). 'Uncertainty and the Welfare Economics of Medical Care'. *American Economic Review*, 53: 941–973.

Ash, A., S. Fienberg, T. Louis, S.-L. T. Normand, T. Stukel and J. Utts (2012). *Statistical issues in assessing hospital performance*. Centre for Medicare & Medicaid Services.

Bailey, T. and P. Hewson (2004). 'Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model'. *Journal of the Royal Statistical Society. Series A*, 167: 501–517.

Bernal-Delgado, E., T. Christiansen, K. Bloor, C. Mateus, A. Yazbeck, J. Munck and J. Bremner (2015). 'ECHO: health care performance assessment in several European health systems'. *European Journal of Public Health*, 25: 3–7.

Berwick, D. (2008). 'The science of improvement'. *Journal of the American Medical Association*, 299: 1182–1184.

Besley, T. and M. Ghatak (2003). 'Incentives, Choice and Accountability in the Provision of Public Services'. *Oxford Review of Economic Policy*, 19: 235–249.

Boadway, R. and N. Bruce (1984). *Welfare economics*. Oxford: Blackwell.

Bradley, E., L. Curry, S. Ramanadhan, L. Rowe, I. Nembhard and H. Krumholz (2009). 'Research in action: using positive deviance to improve quality of health care'. *Implementation Science*, 4: 25.

Briggs, A. and P. Fenn (1998). 'Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane'. *Health Economics*, 7: 723–740.

Brooks, R. (1996). 'EuroQol: the current state of play'. *Health Policy*, 37: 53–72.

Browne, J., L. Jamieson, J. Lewsey, J. van der Meulen, L. Copley and N. Black (2008). 'Case-mix & patients' reports of outcome in Independent Sector Treatment Centres: Comparison with NHS providers'. *BMC Health Services Research*, 8: 78.

Busse, R., J. Schreyögg and P. C. Smith (2008). 'Variability in healthcare treatment costs amongst nine EU countries - results from the HealthBASKET project'. *Health Economics*, 17: S1–S8.

Chard, J., M. Kuczawski, N. Black and J. van der Meulen (2011). 'Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery'. *British Medical Journal*, 343: d6404.

Chua, C., A. Palangkaraya and J. Yong (2010). 'A two-stage estimation of hospital quality using mortality outcome measures: an application using hospital administrative data'. *Health Economics*, 19: 1404–1424.

Coronini-Cronberg, S., J. Appleby and J. Thompson (2013). 'Application of patient-reported outcome measures (PROMs) data to estimate cost-effectiveness of hernia surgery in England'. *Journal of the Royal Society of Medicine*, 106: 278–287.

Daniels, M. and S.-L. T. Normand (2006). 'Longitudinal profiling of health care units based on continuous and discrete patient outcomes'. *Biostatistics*, 7: 1–15.

Dawson, J., R. Fitzpatrick, A. Carr and D. Murray (1996). 'Questionnaire on the perceptions of patients about total hip replacement'. *Journal of Bone & Joint Surgery, British Volume*, 78-B: 185–190.

Department of Health (2008). *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)*. The Stationary Office, London.

— (2012). *Payment by Results Guidance for 2012-13*. The Stationary Office, London.

Devlin, N. J. and J. Sussex (2011). *Incorporating multiple criteria in HTA: methods and processes*. Office for Health Economics, London.

Devlin, N., D. Parkin and J. Browne (2010). 'Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data'. *Health Economics*, 19: 886–905.

Dixit, A. (2002). 'Incentives and Organizations in the Public Sector: An Interpretative Review'. *The Journal of Human Resources*, 37: 696–727.

Dowd, B., T. Swenson, R. Kane, S. Parashuram and R. Coulam (2014). 'Can data envelopment analysis provide a scalar index of 'value'?' *Health Economics*, 23: 1465–1480.

Elixhauser, A., C. Steiner, D. Harris and R. Coffey (1998). 'Comorbidity measures for use with administrative data'. *Medical Care*, 36: 8–27.

Geweke, J., G. Gowrisankaran and R. J. Town (2003). 'Bayesian Inference for Hospital Quality in a Selection Model'. *Econometrica*, 71: 1215–1238.

Glazer, J., T. McGuire and S.-L. T. Normand (2008). 'Mitigating the Problem of Unmeasured Outcomes in Quality Reports'. *The B.E. Journal of Economic Analysis & Policy*, 8: Article 7.

Goddard, M. and R. Jacobs (2009). 'Using composite indicators to measure performance in health care'. In: *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Ed. by P. Smith, E. Mossialos, I. Papanicolas and S. Leatherman. Cambridge: Cambridge University Press. Chap. 3.4, 339–368.

Goddard, M., R. Mannion and P. C. Smith (2000). 'Enhancing performance in health care: a theoretical perspective on agency and the role of information'. *Health Economics*, 9: 95–107.

Goldstein, H. (1986). 'Multilevel mixed linear model analysis using iterative generalized least squares'. *Biometrika*, 73: 43–56.

— (1997). 'Methods in School Effectiveness Research'. *School Effectiveness and School Improvement*, 8: 369–395.

Goldstein, H. and D. J. Spiegelhalter (1996). 'League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance'. *Journal of the Royal Statistical Society: Series A*, 159: 385–443.

Gowrisankaran, G. and R. J. Town (1999). 'Estimating the quality of care in hospitals using instrumental variables'. *Journal of Health Economics*, 18: 747–767.

Gutacker, N., C. Bojke, S. Daidone, N. Devlin and A. Street (2013). 'Hospital Variation in Patient-Reported Outcomes at the Level of EQ-5D Dimensions: Evidence from England'. *Medical Decision Making*, 33: 804–818.

Gutacker, N., A. Street, M. Gomes and C. Bojke (2015). 'Should English healthcare providers be penalised for failing to collect patient-reported outcome measures (PROMs)?' *Journal of the Royal Society of Medicine*, 108: 304–316.

Häkkinen, U., G. Rosenqvist, M. Peltola, S. Kapiainen, H. Rättö, F. Cots, A. Geissler, Z. Or, L. Serdén and R. Sund (2014). 'Quality, cost, and their trade-off in treating AMI and stroke patients in European hospitals'. *Health Policy*, 117: 15–27.

Hall, B. and B. Hamilton (2004). 'New information technology systems and a Bayesian hierarchical bivariate probit model for profiling surgeon quality at a large hospital'. *The Quarterly Review of Economics and Finance*, 44: 410–429.

Hauck, K., N. Rice and P. C. Smith (2003). 'The influence of health care organisations on health system performance'. *Journal of Health Services Research & Policy*, 8: 68–74.

Hauck, K. and A. Street (2006). 'Performance assessment in the context of multiple objectives: A multivariate multilevel analysis'. *Journal of Health Economics*, 25: 1029–1048.

Holmström, B. and P. Milgrom (1991). 'Multi-task principle-agent problems: Incentive contracts, asset owernship and job design'. *Journal of Law, Economics and Organization*, 7: 24–52.

Hussey, O., S. Wertheimer and A. Mehrotra (2013). 'The Association Between Health Care Quality and Cost: A Systematic Review'. *Annals of Internal Medicine*, 158: 27–34.

Husted, H., G. Holm and S. Jacobsen (2008). 'Predictors of length of stay and patient satisfaction after hip and knee replacement surgery - Fast-track experience in 712 patients'. *Acta Orthopaedica*, 79: 168–173.

Karnon, J., O. Caffrey, C. Pham, R. Grieve, D. Ben-Tovim, P. Hakendorf and M. Crotty (2013). 'Applying risk adjusted cost-effectiveness (RAC-E) analysis to hospitals: Estimating the costs and consequences of variation in clinical practice'. *Health Economics*, 22: 631–642.

Keeler, E. B. (1990). 'What proportion of hospital cost differences is justifiable?' *Journal of Health Economics*, 9: 359–65.

Kruse, M. and J. Christensen (2013). 'Is quality costly? Patient and hospital cost drivers in vascular surgery'. *Health Economics Review*, 3: 22.

Landrum, M., S. Bronskill and S.-L. T. Normand (2000). 'Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers'. *Health Services and Outcomes Research Methodology*, 1: 23–47.

Landrum, M., S.-L. T. Normand and R. Rosenheck (2003). 'Selection of Related Multivariate Means'. *Journal of the American Statistical Association*, 98: 7–16.

Larsen, K., O. Sørensen, T. Hansen, P. Thomsen and K. Søballe (2008). 'Accelerated perioperative care and rehabilitation intervention for hip and knee replacement is effective: A randomized clinical trial involving 87 patients with 3 months of follow-up'. *Acta Orthopaedica*, 79: 149–159.

Lawton, R., N. Taylor, R. Clay-Williams and J. Braithwaite (2014). 'Positive deviance: a different approach to achieving patient safety'. *BMJ Quality & Safety*, 23: 880–883.

Leckie, G. and C. Charlton (2013). 'runmlwin: Stata module for fitting multilevel models in the MLwiN software package'. *Journal of Statistical Software*, 52: 1–40.

Little, R. and D. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Martin, S. and P. C. Smith (2005). 'Multiple Public Service Performance Indicators: Toward an Integrated Statistical Approach'. *Journal of Public Administration Research and Theory*, 15: 599–613.

McGuire, A. (2001). 'Theoretical concepts in the economic evaluation of health care'. In: *Economic evaluation in health care - merging theory and practice*. Ed. by M. Drummond and A. McGuire. Oxford University Press.

Monitor and NHS England (2013). *National Tariff Payment System - Annex 4A: Additional information on currencies with national prices*.

Mundlak, Y. (1978). 'On the Pooling of Time Series and Cross Section Data'. *Econometrica*, 46: 69–85.

National Clinical Audit Advisory Group (2011). *Detection and management of outliers*. The Stationary Office, London.

Neuburger, J., A. Hutchings, J. van der Meulen and N. Black (2013). 'Using patient-reported outcomes (PROs) to compare the provider of surgery: does the choice of measure matter?' *Medical Care*, 51: 517–523.

Noble, M., G. Wright, G. Smith and C. Dibben (2006). 'Measuring multiple deprivation at the small-area level'. *Environment and Planning A*, 38: 169–185.

Normand, S.-L. T., M. Glickman and C. Gatsonis (1997). 'Statistical Methods for Profiling Providers of Medical Care: Issues and Applications'. *Journal of the American Statistical Association*, 92: 803–814.

O'Hagan, A., J. Stevens and J. Montmartin (2000). 'Inference for the Cost-Effectiveness Acceptability Curve and Cost-Effectiveness Ratio'. *PharmacoEconomics*, 17: 339–349.

Paton, F., D. Chambers, P. Wilson, A. Eastwood, D. Craig, D. Fox, D. Jayne and E. McGinnes (2014). 'Effectiveness and implementation of enhanced recovery after surgery programmes: a rapid evidence synthesis'. *BMJ Open*, 4: e005015.

Pedraja-Chaparro, F., J. Salinas-Jimenez and P. Smith (1999). 'On the Quality of the Data Envelopment Analysis Model'. *The Journal of the Operational Research Society*, 50: 636–644.

Porter, M. E. (2010). 'What Is Value in Health Care?' *The New England Journal of Medicine*, 363: 2477–2481.

Portrait, F., O. Galiën and B. van den Berg (2015). 'Measuring healthcare providers' performance within managed competition using multidimensional quality and cost indicators'. *Health Economics*, forthcoming. DOI: 10.1002/hec.3158.

Propper, C., M. Sutton, C. Whitnall and F. Windmeijer (2008). 'Did 'Targets and Terror' Reduce Waiting Times in England for Hospital Care?' *The B.E. Journal of Economic Analysis & Policy*, 8: Article 5.

Propper, C. and D. Wilson (2012). 'The use of performance measures in health care systems'. In: *The Elgar Companion to Health Economics*. Ed. by A. M. Jones. 2nd ed. Edward Elgar. Chap. 33, 350–358.

Ryan, M., D. Scott, C. Reeves, A. Bate, E. van Teijlingen, E. Russell, M. Napper and C. Robb (2001). 'Eliciting public preferences for healthcare: a systematic review of techniques'. *Health Technology Assessment*, 5: 1–186.

Shleifer, A. (1985). 'A Theory of Yardstick Competition'. *RAND Journal of Economics*, 16: 319–327.

Siciliani, L., P. Sivey and A. Street (2013). 'Differences in length of stay for hip replacement between public hospital, specialised treatment centres and private providers: selection or efficiency?' *Health Economics*, 22: 234–242.

Skrondal, A. and S. Rabe-Hesketh (2009). 'Prediction in multilevel generalized linear models'. *Journal of the Royal Statistical Society: Series A*, 172: 659–687.

Smith, P. C. and A. Street (2005). 'Measuring the efficiency of public services: the limits of analysis'. *Journal of the Royal Statistical Society. Series A*, 168: 401–417.

Smith, P. C. (2002). 'Developing composite indicators for assessing health system efficiency'. In: *Measuring up - Improving health system performance in OECD countries*. Ed. by OECD. OECD Publications Service. Chap. 14, 295–316.

Spiegelhalter, D. J. (2005). 'Funnel plots for comparing institutional performance'. *Statistics in Medicine*, 24: 1185–1202.

Street, A., N. Gutacker, C. Bojke, N. Devlin and S. Daidone (2014). 'Variation in outcome and costs among NHS providers for common surgical procedures: econometric analysis of routinely collected data'. *Health Services and Delivery Research*, 2: 1–89.

Teixeira-Pinto, A. and S.-L. T. Normand (2008). 'Statistical methodology for classifying units on the basis of multiple-related measures'. *Statistics in Medicine*, 27: 1329–1350.

Terza, J., A. Basu and P. Rathouz (2008). 'Two-stage residual inclusion estimation: Addressing endogeniety in health econometric modeling'. *Journal of Health Economics*, 27: 531–543.

Thum, Y. (1997). 'Hierarchical Linear Models for Multivariate Outcomes'. *Journal of Educational and Behavioral Statistics*, 22: 77–108.

Timbie, J. W., J. P. Newhouse, M. B. Rosenthal and S.-L. T. Normand (2008). 'A Cost-Effectiveness Framework for Profiling the Value of Hospital Care'. *Medical Decision Making*, 28: 419–434.

Timbie, J. W. and S.-L. T. Normand (2008). 'A comparison of method for combining quality and efficiency performance measures: Profiling the value of hospital care following acute myocardial infarction'. *Statistics in Medicine*, 27: 1351–1370.

Timbie, J. W., D. Shahian, J. Newhouse, M. B. Rosenthal and S.-L. T. Normand (2009). 'Composite measures for hospital quality using quality-adjusted life years'. *Statistics in Medicine*, 28: 1238–1254.

Van Hout, B., M. Al, G. Gordon and F. Rutten (1994). 'Costs, effects, and C/E-ratios alongside a clinical trial'. *Health Economics*, 3: 309–319.

Wennberg, J. and A. Gittelsohn (1973). 'Small Area Variation in Health Care Delivery: A population-based health information system can guide planning and regulatory decision-making'. *Science*, 182: 1102–1108.

Zellner, A. (1962). ‘An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias’. *Journal of the American Statistical Association*, 57: 348–368.

7. Appendix

Table A1: Estimated coefficients and standard errors from multivariate regression model

	Length of stay		Post-operative OHS		Waiting time $>$ 18 weeks		28-day emergency readmission	
Variable	Est	SE	Est	SE	Est	SE	Est	SE
Constant	2.078	0.052***	27.154	0.823***	-1.335	0.053***	-1.609	0.119***
FY 2010/11	-0.096	0.011***	0.043	0.072	0.114	0.045*	-0.008	0.019
FY 2011/12	-0.203	0.015***	0.229	0.085**	0.208	0.050***	-0.051	0.020*
Pre-operative OHS	-0.011	0.001***	0.599	0.016***			-0.005	0.001***
Pre-operative OHS^2	0.000	0.000***	-0.009	0.000***				
Patient age	-0.027	0.002***	0.208	0.025***			-0.014	0.004***
Patient age^2	0.000	0.000***	-0.002	0.000***			0.000	0.000***
Male patient	-0.074	0.004***	0.908	0.062***			0.142	0.015***
Primary diagnosis: Rheumatoid arthritis	0.026	0.022	-0.486	0.529			-0.086	0.113
Primary diagnosis: Other	0.035	0.009***	-1.169	0.187***			0.079	0.028**
Elixhauser comorbidities: 1	0.025	0.004***	-0.456	0.068***			0.061	0.017***
Elixhauser comorbidities: 2-3	0.068	0.004***	-1.433	0.083***			0.148	0.017***
Elixhauser comorbidities: 4+	0.153	0.007***	-2.826	0.133***			0.285	0.023***
Previously admitted as an emergency	0.071	0.005***	-0.613	0.124***			0.137	0.023***
Socio-economic status	0.003	0.001**	-0.523	0.027***			0.011	0.005*
Disabled	-0.036	0.003***	2.586	0.080***			-0.065	0.016***
Living alone	0.111	0.005***	-0.368	0.071***				
Symptom duration: 1 - 5 years	0.020	0.004***	-0.654	0.077***				
Symptom duration: 6 - 10 years	0.039	0.005***	-1.335	0.121***				
Symptom duration: $>$ 10 years	0.055	0.007***	-1.712	0.159***				
Assistance in filling in PROM questionnaire	0.067	0.005***	-0.545	0.097***				
HRG: HB11C - category 1 without CC	0.037	0.006***						
HRG: HB12B - category 2 with CC	0.127	0.006***						
HRG: HB12A - category 2 with major CC	0.495	0.011***						
HRG: HB11B - category 1 with CC	0.122	0.016***						
HRG: other	0.376	0.031***						
First-stage residual	0.001	0.000	0.012	0.006*	0.001	0.001	0.003	0.001*
Var(θ_j)	0.025	0.002***	1.203	0.141***	0.378	0.038***	0.023	0.003***
Var(ϵ_{ij})	0.162	0.001***	68.563	0.340***	1.000		1.000	
Number of observations	95,955							

*** p$<$ 0.001; ** p$<$0.01; * p$<$0.05

Legend: Est = Estimate; SE = Huber-White standard error (robust to unknown heteroscedasticity); OHS = Oxford Hip Score; HRG = Healthcare Resource Group; CC = Complications and comorbidities; FY = Financial year (April - March). Socio-economic status is approximated by the % of neighbourhood residents claiming income benefits. This characteristics is measured at neighbourhood level (lower super output area (LSOA)).

Table A2: Estimated coefficients and standard errors - multinomial hospital choice model (first-stage)

Variable	Est	SE
Closest hospital	0.185	0.014***
Distance to hospital	-0.197	0.003***
Distance^2	0.001	0.0001***
Distance^3	-0.00002	0.000002***
Number of patients	95,955	
Number of providers	252	
Pseudo R^2	0.706	
$\chi^2(4)$	120,930	

*** p$<$ 0.001; ** p$<$0.01; * p$<$0.05
Legend: Est = Estimate; SE = Huber-White standard error
Notes: Distance to hospital is measured as the straight-line distance from the centroid of the patient's lower super output area (LSOA) to the provider's headquarter (NHS trust) or hospital site (ISTCs). Distance is measured in kilometres.

Table A3: Correlation between performance dimensions - excluding ISTCs

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.27**	1.00	***-0.02***	***-0.07***
Waiting time $>$ 18 wks (3)	0.11	-0.02	1.00	*0.00*
28-day emergency readmission (4)	-0.03	**-0.46**	-0.02	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

Table A4: Correlation between performance dimensions - accounting for provider average risk factors

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.13***	***0.02***	***0.02***
Post-operative OHS (2)	**-0.21**	1.00	***-0.02***	***-0.07***
Waiting time $>$ 18 wks (3)	**0.19**	**-0.17**	1.00	*0.00*
28-day emergency readmission (4)	-0.08	**-0.35**	0.07	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.

Table A5: Correlation between performance dimensions - risk-adjustment based on HES data only

Performance dimension	(1)	(2)	(3)	(4)
Length of stay (1)	1.00	***-0.16***	***0.01***	***0.02***
Post-operative OHS (2)	**-0.41**	1.00	***-0.01***	***-0.07***
Waiting time $>$ 18 wks (3)	**0.28**	**-0.37**	1.00	*0.00*
28-day emergency readmission (4)	0.04	**-0.47**	**0.17**	1.00

Notes: Lower triangle reports the correlation between random effects at provider level, whereas upper triangle (in italics) reports the correlation between random effects (i.e. the idiosyncratic error term) at patient level. Bold indicates that the correlation is statistically significantly different from zero at the 95% level.